ALE Analytics - A microbial data warehouse and analytical suite

Patrick V. Phaneuf[1], Dennis Gosting[2], Bernhard O. Palsson[2,3], Adam M. Feist[2,3]

pphaneuf@eng.ucsd.edu, afeist@ucsd.edu
1) Department of Computer Science and Engineering, University of California at San Diego, La Jolla 92093-0404, United States of America,
2) Department of Bioengineering, University of California San Diego, La Jolla, CA, 92093, United States of America,
3) Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800, Lyngby, Denmark

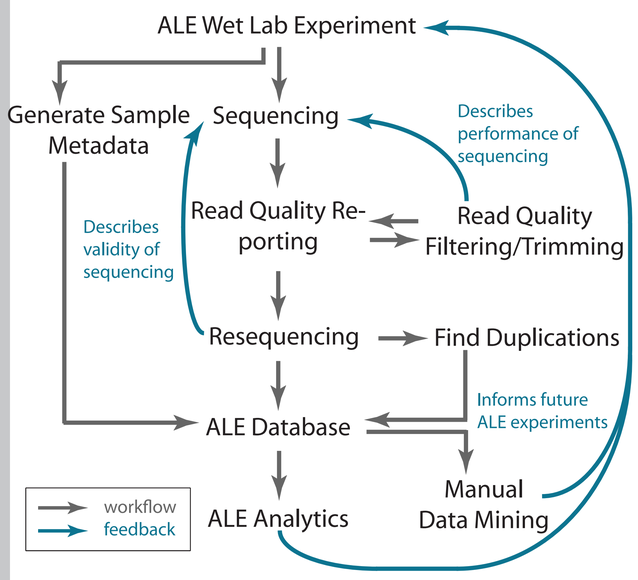
systems biology research group

Abstract

Adaptive Laboratory Evolution (ALE) is a powerful experimental tool used by academic and industrial labs for a number of applications. ALE experiments generate a substantial amount of data, coming in the form of sequencing reads, alignment reports, and sample metadata. As ALE experiments scale to include more samples, the task of managing this data comes at higher costs due to the effort necessary to organize and integrate data into a format that describes the evolution process succinctly. In this work, we describe the development, deployment and iteration of an 'ALE Analytics' pipeline and web platform that streamlines the necessary ALE experiment data post-processing, manages experiment data, and produces interactive reports that detail an ALE experiment. Our design has been primarily driven by the need to consolidate large amounts of ALE experimental data in such a way to describe the quality of the sample sequencing, adaptive mutations in evolved strains, the context of mutations via their metadata (i.e., culturing environments, strain properties), and related mutations found in other experiments housed in the database. We have done so by leveraging a full stack of technologies that enable the parsing and databasing of experiment data, the execution of automated analysis on said data and the generation of web accessible reports. Future efforts will take full advantage of this developing platform to enable more depth and breadth of ALE experiment analysis with quicker turnaround.

ALE Analytics Pipeline

Our ALE pipeline leverages industry established tools to process and refine ALE replicate sequencing data to identify genomic perturbations and integrate them into a database that can be mined for mutational trends.



Features and Advantages

- Supports multiple bacterial strains.
- Leverages established tools:
  Breseq[3] (resequencing),
  FastQC[4] (read quality reporting),
  FASTX-Toolkit[5] (read trimming)
- Automated multi-sample processing for specific stages.
- Stage dependent feedback.
- Feedback provides opportunity for experiment and data refinement.

ALE Analytics Web Platform Features

The ALE Analytics web platform integrates our ALE mutation analysis with published multi-omics data and better enables identification of causal mutations in ALE experiments.

ALE Database

Database containing the mutational data and metadata generated by the ALE data processing pipeline. This database is the basis for all analytical features implemented within the ALE Analytics web platform

ALE Analytics

A web interface reporting ALE experiment data and trends found through automated datamining.

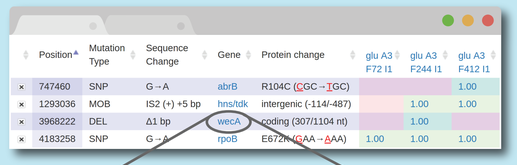
Sample Metadata

Contextualizing mutations using descriptions of each ALE sample: strain, genetic perturbations, media, substrate, temperature, etc.

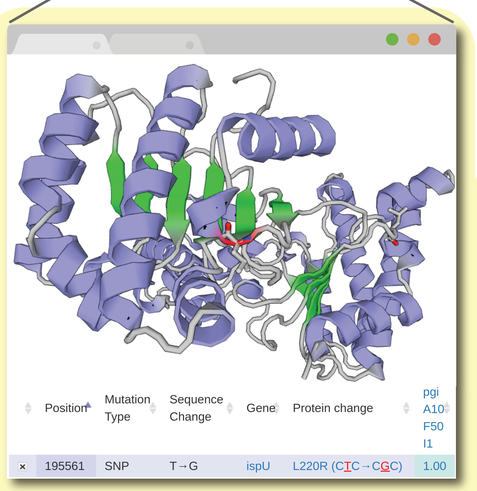
ALE Experiment trend reporting

Mutation Lineages

A report of ALE samples and their mutations. Fundamental feature used in understanding an ALE's mutational evolution, where one visually inspects for mutational trends, such as fixation, among all samples within an ALE experiment.[1]



Position	Mutation Type	Sequence Change	Gene	Protein change	glu A3 F72 I1	glu A3 F244 I1	glu A3 F412 I1
× 747460	SNP	G→A	abrB	R104C (CGC→TGC)			1.00
× 1293036	MOB	IS2 (+) +5 bp	hns/tdk	intergenic (-114/-487)		1.00	1.00
× 3968222	DEL	Δ1 bp	wecA	coding (307/1104 nt)		1.00	
× 4183258	SNP	G→A	rpoB	E672K (GAA→AAA)	1.00	1.00	1.00

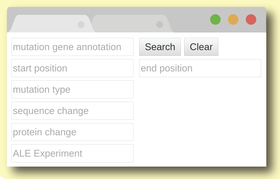


Position	Mutation Type	Sequence Change	Gene	Protein change	pgi A10 F50 I1
× 195561	SNP	T→G	ispU	L220R (CTC→CGC)	1.00

ALE Database trend reporting

Mutation Search

Search the ALE Database by mutation specific features.



Genes

Gene specific information:
- All ALE Database mutations of a particular gene.
- 3D rendering[6] of a gene's product.
- Highlight structure of a gene's product effected by a mutation.
- Display's ligand structure at site of mutation

Common Mutations

A report of ALE samples and their sets of shared mutations. Used to find samples most mutationally similar to a currently selected sample.

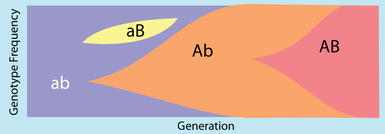
Frequently Mutated Genes

A report of an ALE experiment's multiple replicates and their sets of mutations that:
1. Affect the same gene within a replicate.
2. Affect the same gene among parallel replicates.

Highlights which genes are more frequently mutated within an ALE experiment, potentially elucidating the functional categories that were more causal.

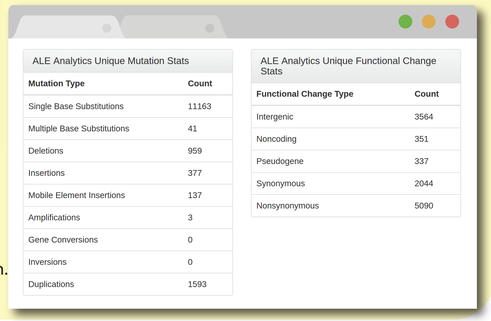
Mutation Fixation

A report of ALE samples and their mutations that emerged and fixed within an ALE. Presents which mutations were kept by the dominant populations between flasks and therefore may have lead to fitness benefits.



Dashboard

General statistics on entire ALE Database including:
- Most frequently mutated genes
- Most frequent mutations
- Mutation type counts (SNP, INDEL, etc.)



ALE Analytics Unique Mutation Stats

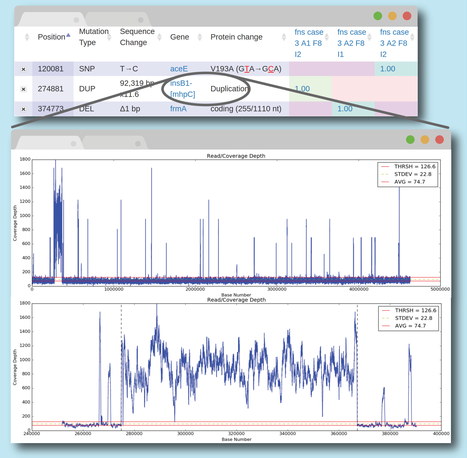
Mutation Type	Count
Single Base Substitutions	11163
Multiple Base Substitutions	41
Deletions	959
Insertions	377
Mobile Element Insertions	137
Amplifications	3
Gene Conversions	0
Inversions	0
Duplications	1593

ALE Analytics Unique Functional Change Stats

Functional Change Type	Count
Intergenic	3564
Noncoding	351
Pseudogene	337
Synonymous	2044
Nonsynonymous	5090

Duplications

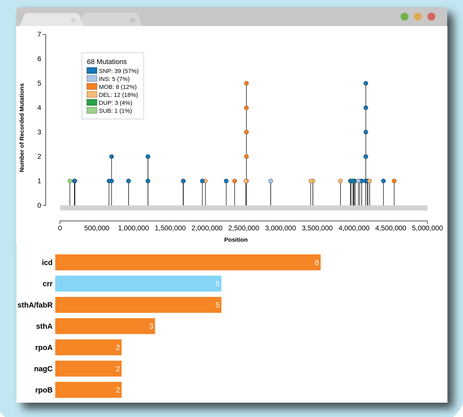
Genetic duplication and amplification identification based on resequencing alignment read-depth. Can find large areas of duplication that are not otherwise found with current resequencing tools.[2]



Position	Mutation Type	Sequence Change	Gene	Protein change	fns case 3 A1 F8 I2	fns case 3 A2 F8 I1	fns case 3 A2 F8 I2
× 120081	SNP	T→C	aceE	V193A (GTA→GCA)			1.00
× 274881	DUP	92,319 bp ↔11.6	[insB1-]mhpC]	Duplication	1.00		
× 374773	DEL	Δ1 bp	frmA	coding (255/1110 nt)		1.00	

Experiment Statistics

Statistics and visualizations that can lend intutition to an ALE experiment's sequencing data quality, mutational "hotspots"[7], most frequently mutated genes and most frequent mutations.

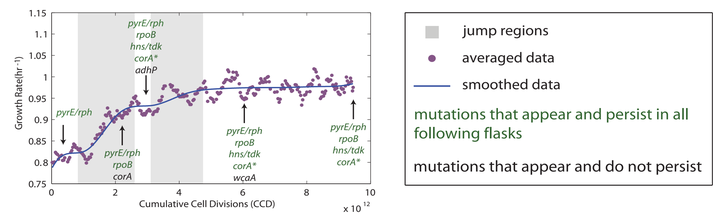


References

1) LaCroix RA et al. Appl Environ Microbiol 2014, 81:17-30
2) Guzman GI et al. PNAS 2015, 112: 929-934
3) Deatherage DE et al. Methods Mol. Biol. 1151: 165–188
4) Andrews S. (2010). FastQC: a quality control tool for high throughput sequencedata. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
5) The FASTX-Toolkit. Available online at: http://hannonlab.cshl.edu/fastx_toolkit/
6) DOI 10.5281/zenodo.20980
7) DOI 10.5281/zenodo.14561

Future Directions

- Statistical analysis to rank novelty of mutations including functional clustering.
- Integrate mutations with structural annotations for structural systems biology.
- Annotation for regulatory mutations though omics data integration (RNAseq).
- Make a public branch for platform and database (currently internal only).
- Large-scale analysis of evolution dynamics from many projects and samples.
- Multiple dimension ranking for project-specific and strain-specific mutations.
- Datamine for pairwise mutations shared among different ALE Experiments and establish their relationship to a population.



- Integrate ALE experiment fitness[1] data to provide additional context for identifying causal mutations. As shown, mutations that appear during an exponential growth phase are more likely correlated to fitness benefits